Data Preprocessing and Exploratory Data Analysis (EDA)

1. Auto MPG

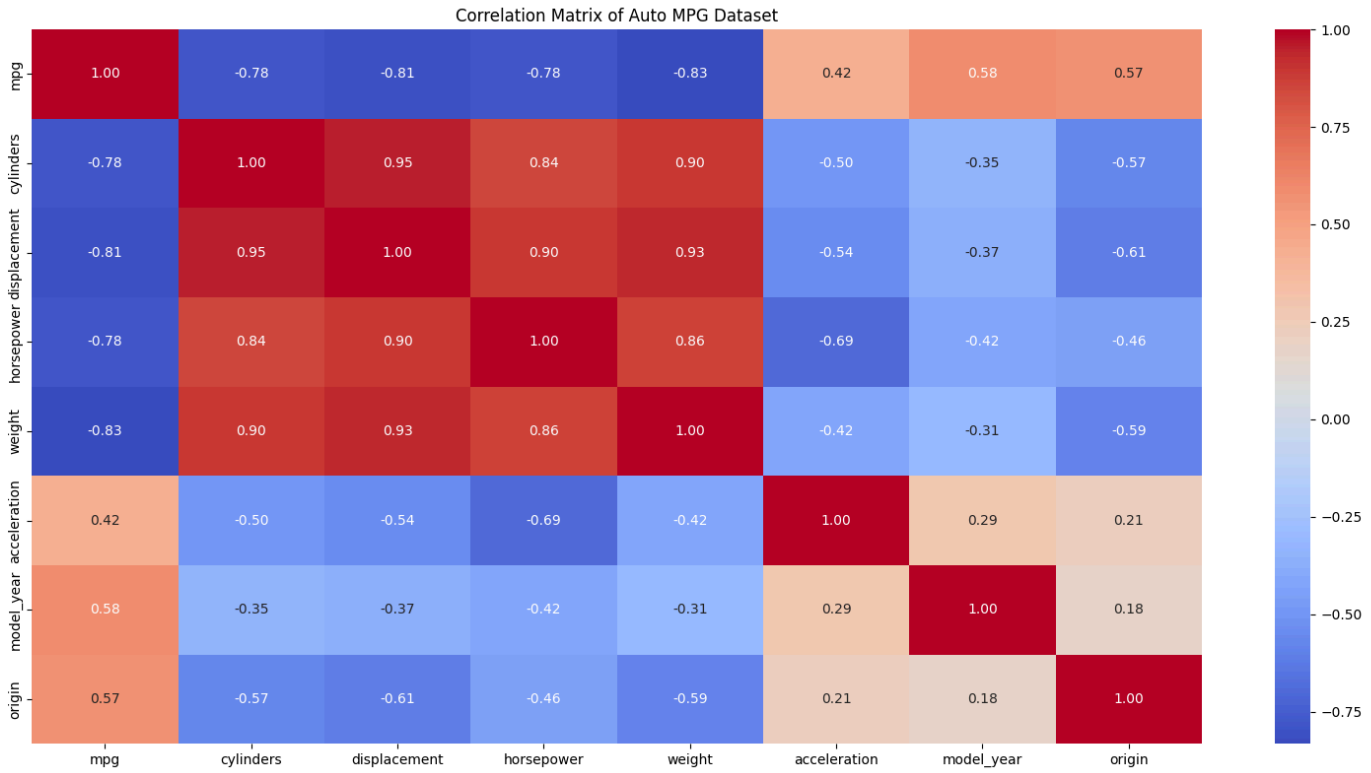
Datatypes

Name	Type	Name	Type
mpg	float64	acceleration	float64
cylinders	int64	model_year	int64
displacement	float64	origin	int64
horsepower	float64	car_name	object
weight	float64		

Missing Data and Shape

Shape with n/a values: (398, 9)
Shape without n/a values: (392, 9)
n/a values: 6

Correlation Matrix Heatmap



Group Analysis

Average MPG by Brand

Brand	MPG	Brand	MPG	Brand	MPG	Brand	MPG
amc	18.070370	chevrolet	20.370455	ford	19.475000	mercury	19.118182
audi	26.714286	chevy	18.000000	hi	9.000000	nissan	36.000000
bmw	23.750000	chrysler	17.266667	honda	33.761538	oldsmobile	21.100000
buick	19.182353	datsun	31.113043	mazda	30.058333	opel	25.750000
cadillac	19.750000	dodge	22.060714	mercedes	25.400000	peugeot	23.687500
capri	25.000000	fiat	28.912500	volkswagen	29.150000	plymouth	21.703226
pontiac	20.012500	subaru	30.525000	toyota	28.165385	vw	39.016667
renault	29.666667	saab	23.900000	triumph	35.000000	volvo	21.116667
Mercedes-benz	23.250000						

Average MPG by cylinders

Cylinders	MPG	Cylinders	MPG	Cylinders	MPG
3	20.550000	4	29.283920	5	27.366667
6	19.973494	8	14.963107		

Average MPG by model year

model year	mpg	model year	mpg	model year	mpg
70	17.689655	75	20.266667	79	25.093103
71	21.111111	76	21.573529	80	33.803704
72	18.714286	77	23.375	81	30.185714
73	17.1	78	24.061111	82	32
74	22.769231				

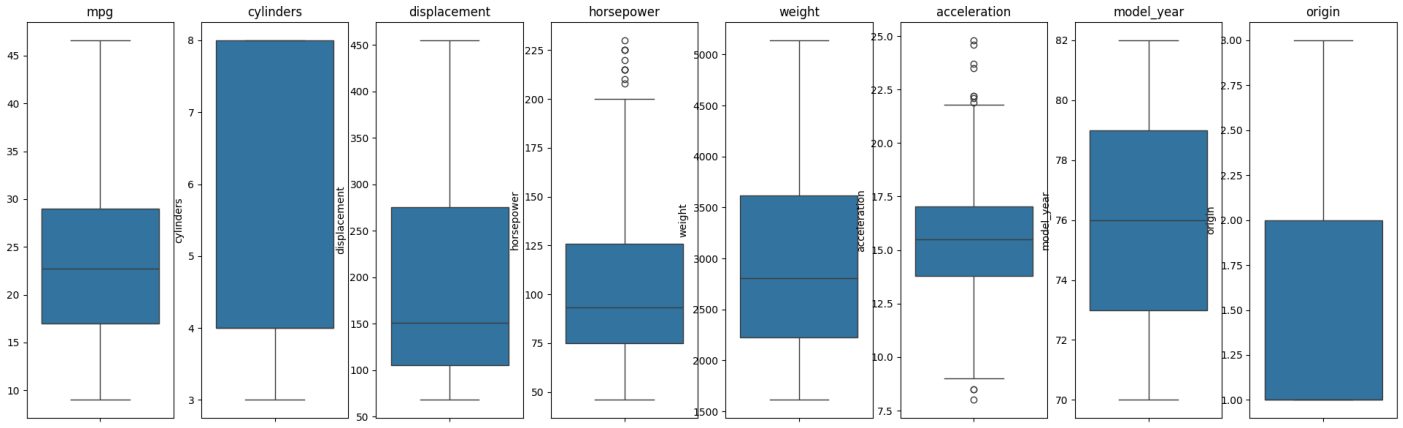
Average MPG by origin

origin	mpg	origin	mpg	origin	mpg
1	20.033469	2	27.602941	3	0.450633

Descriptive Statistics

features	count	mean	std	min	25%	50%	75%	max	skewness	kurtosis
mpg	392	23.445918	7.805007	9	17	22.75	29	46.6	0.457092	-0.515993
cylinders	392	5.471939	1.705783	3	4	4	8	8	0.508109	-1.398199
displacement	392	194.41199	104.644004	68	105	151	275.75	455	0.701669	-0.778317
horsepower	392	104.469388	38.49116	46	75	93.5	126	230	1.087326	0.696947
weight	392	2977.584184	849.40256	1613	2225.25	2803.5	3614.75	5140	0.519586	-0.809259
acceleration	392	15.541327	2.758864	8	13.775	15.5	17.025	24.8	0.291587	0.444234
model_year	392	75.979592	3.683737	70	73	76	79	82	0.019688	-1.167446
origin	392	1.576531	0.805518	1	1	1	2	3	0.915185	-0.841885

Distribution Statistics



2. Forest Fire

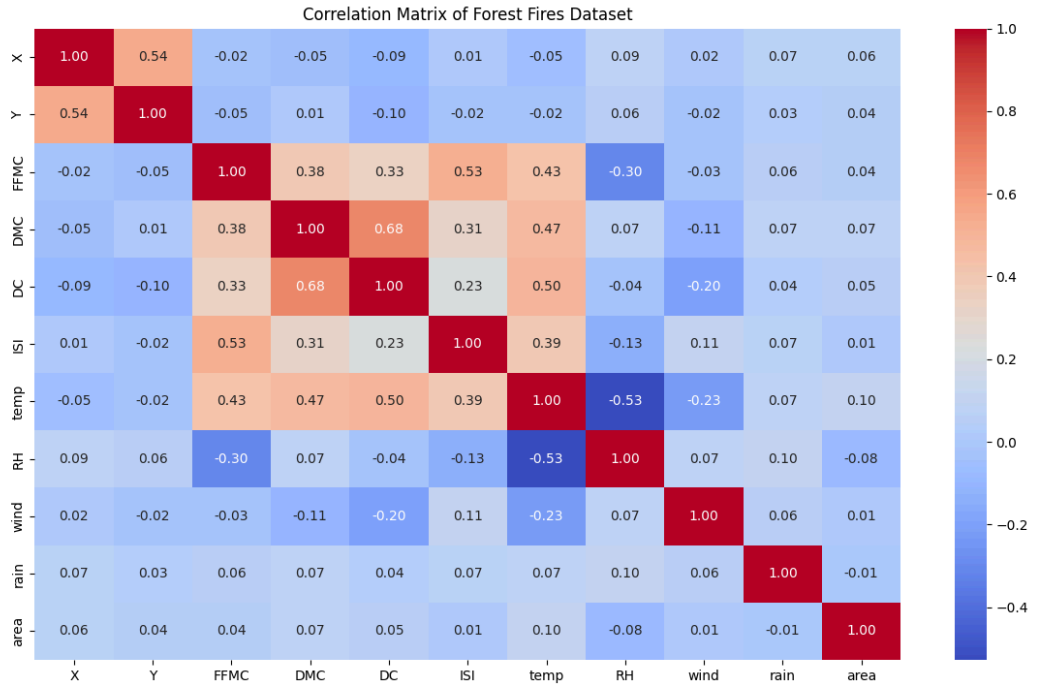
Datatypes

Name	Type	Name	Type
X	int64	ISI	float64
Y	int64	temp	float64
month	object	RH	int64
Day	object	wind	float64
FFMC	float64	rain	float64
DMC	float64	area	float64
DC	float64	dtype	object

Missing Data and Shape

Shape with n/a values: (517, 13)
Shape without n/a values: (517, 13)
n/a values: 0

Correlation Matrix Heatmap



Group Analysis

Average area by month

Month	Area
apr	8.891111
aug	12.489076
dec	13.330000
feb	6.275000
jan	0.000000
jul	14.369687
jun	5.841176
mar	4.356667
may	19.240000
nov	0.000000
oct	6.638000
sep	17.942616

Average area by day

Day	Area
fri	5.261647
mon	9.547703
sat	25.534048
sun	10.104526
thu	16.345902
tue	12.621719
wed	10.714815

Average area by rain

Rain	Area
0.0	13.023694
0.2	0.000000
0.4	0.000000
0.8	0.000000
1.0	0.000000
1.4	2.170000
6.4	10.820000

Average area by X

X	Area
1	13.392292
2	9.570548
3	2.456545
4	10.385165
5	3.045667
6	20.115000
7	11.092667
8	24.466885
9	18.546923

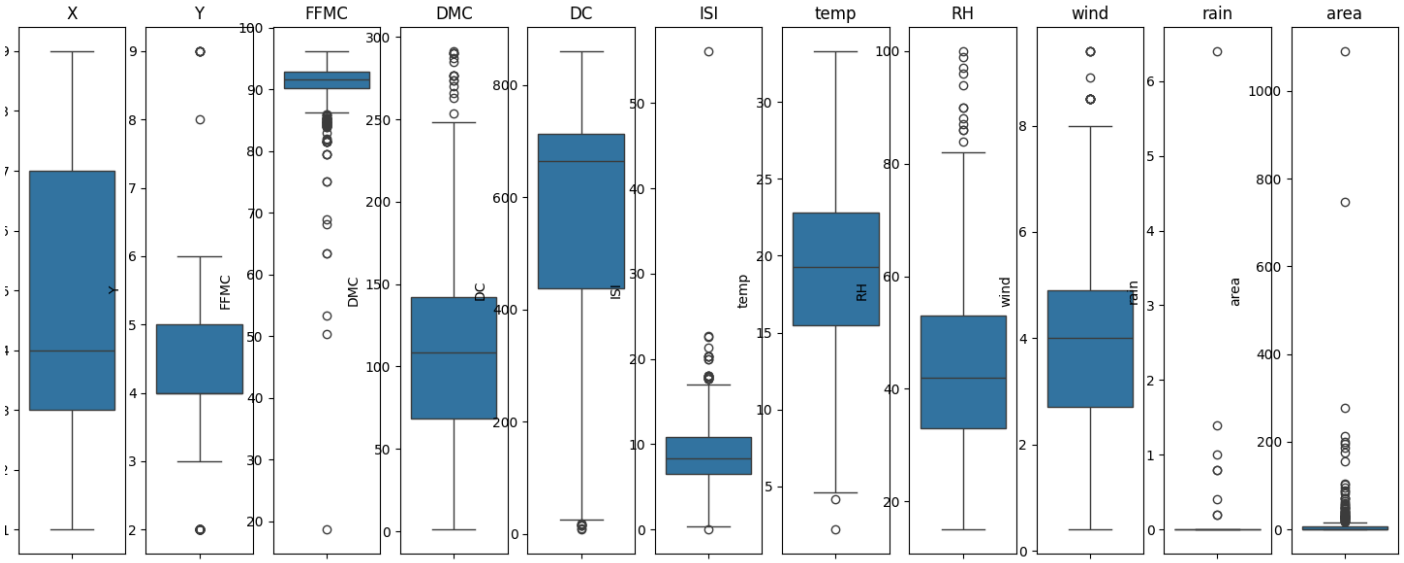
Average area by Y

Y	Area
2	15.513409
3	9.110000
4	8.412857
5	15.758560
6	20.385946
8	185.760000
9	0.745000

Descriptive Statistics

features	count	mean	std	min	25%	50%	75%	max	skewness	kurtosis
X	517.0	4.669246	2.313778	1.0	3.0	4.00	7.00	9.00	0.036246	-1.172331
Y	517.0	4.299807	1.229900	2.0	4.0	4.00	5.00	9.00	0.417296	1.420553
FFMC	517.0	90.644681	5.520111	18.7	90.2	91.60	92.90	96.20	-6.575606	67.066041
DMC	517.0	110.872340	64.046482	1.1	68.6	108.30	142.40	291.30	0.547498	0.204822
DC	517.0	547.940039	248.066192	7.9	437.7	664.20	713.90	860.60	-1.100445	-0.245244
ISI	517.0	9.021663	4.559477	0.0	6.5	8.40	10.80	56.10	2.536325	21.458037
temp	517.0	18.889168	5.806625	2.2	15.5	19.30	22.80	33.30	-0.331172	0.136166
RH	517.0	44.288201	16.317469	15.0	33.0	42.00	53.00	100.00	0.862904	0.438183
wind	517.0	4.017602	1.791653	0.4	2.7	4.00	4.90	9.40	0.571001	0.054324
rain	517.0	0.021663	0.295959	0.0	0.0	0.00	0.00	6.40	19.816344	421.295964
area	517.0	12.847292	63.655818	0.0	0.0	0.52	6.57	1090.84	12.846934	194.140721

Distribution Statistics



3. Seoul Bike Share Demand

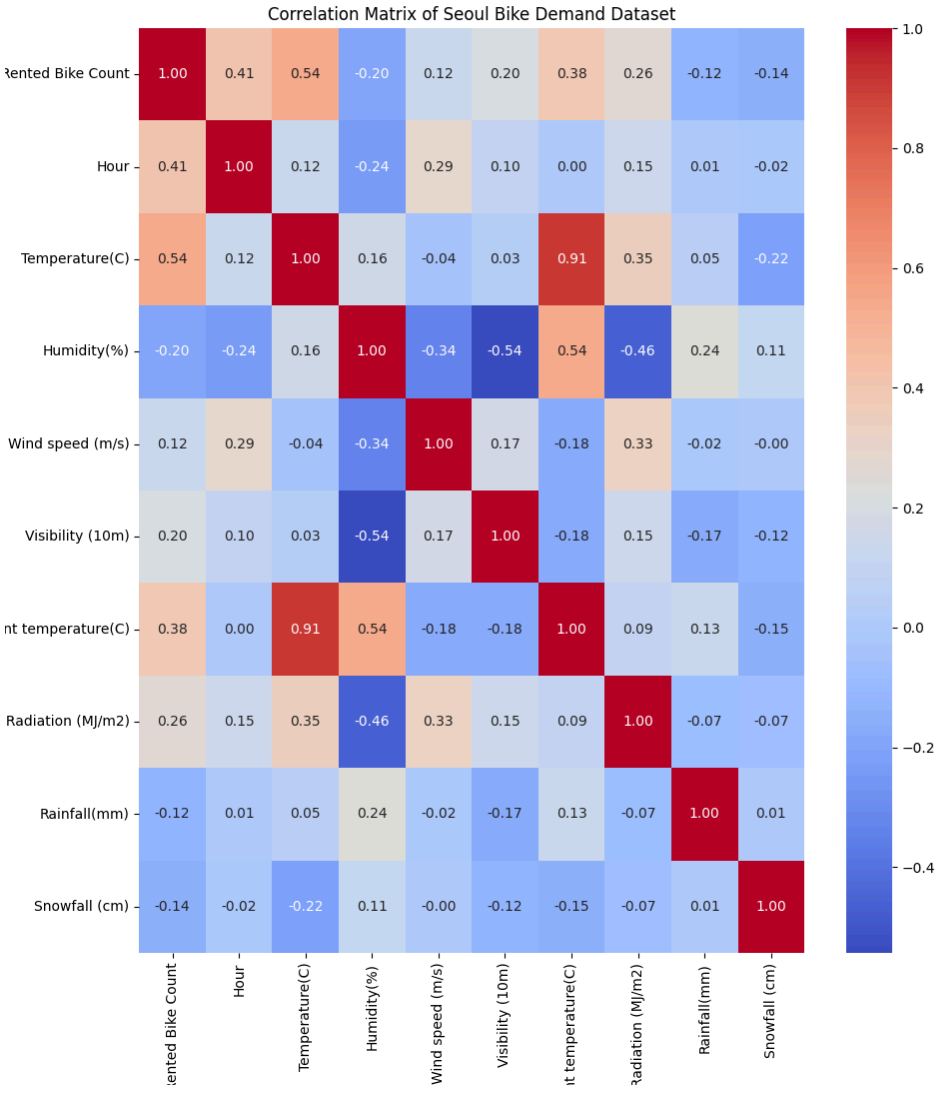
Datatypes

Date	object	Dew point temperature(C)	float64
Rented Bike Count	int64	Solar Radiation (MJ/m2)	float64
Hour	int64	Rainfall(mm)	float64
Temperature(C)	float64	Snowfall (cm)	float64
Humidity(%)	int64	Seasons	object
Wind speed (m/s)	float64	Holiday	object
Visibility (10m)	int64	Functioning Day	object

Missing Data and Shape

Shape with n/a values: (8760, 14)
Shape without n/a values: (8760, 14)
n/a values: 0

Correlation Matrix Heatmap



Group Analysis

Average Rented Bike Count by Seasons

Seasons	Rented Bike Count
Autumn	819.597985
Spring	730.03125
Summer	1034.07337
Winter	225.541204

Average by Holiday

Holiday	Rented Bike Count
Holiday	499.756944
No Holiday	715.228026

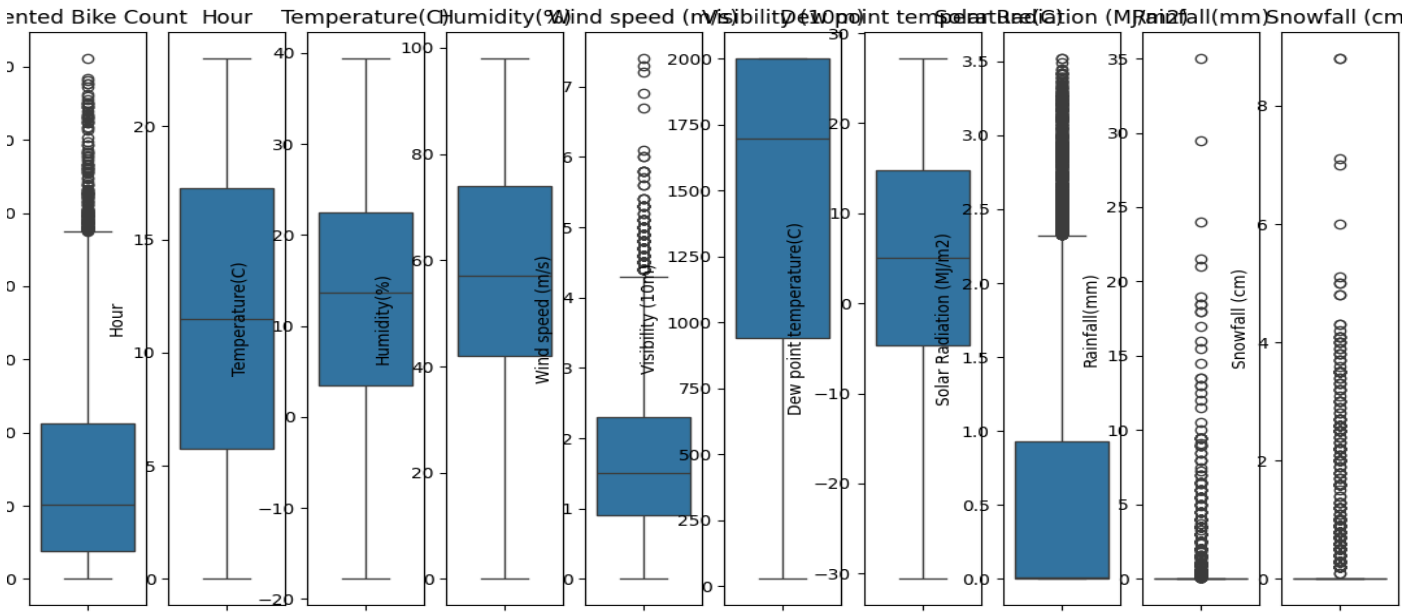
Average by Hours

Hour	Rented Bike Count	Hour	Rented Bike Count
0	541.460274	12	699.441096
1	426.183562	13	733.246575
2	301.630137	14	758.824658
3	203.331507	15	829.186301
4	132.591781	16	930.621918
5	139.082192	17	1138.509589
6	287.564384	18	1502.926027
7	606.005479	19	1195.147945
8	1015.70137	20	1068.964384
9	645.983562	21	1031.449315
10	527.821918	22	922.79726
11	600.852055	23	671.126027

Descriptive Statistics

Metrics	Count	Mean	Std	Min	25%	50%	75%	Max	Skewness	Kurtosis
Rented Bike Count	8760	704.602055	644.997468	0	191	504.5	1065.25	3556	1.153428	0.853387
Hour	8760	11.5	6.922582	0	5.75	11.5	17.25	23	0	-1.204176
Temperature(C)	8760	12.882922	11.944825	-17.8	3.5	13.7	22.5	39.4	-0.198326	-0.837786
Humidity(%)	8760	58.226256	20.362413	0	42	57	74	98	0.059579	-0.803559
Wind speed (m/s)	8760	1.724909	1.0363	0	0.9	1.5	2.3	7.4	0.890955	0.727179
Visibility (10m)	8760	1436.825799	608.298712	27	940	1698	2000	2000	-0.701786	-0.96198
Dew point temperature(C)	8760	4.073813	13.060369	-30.6	-4.7	5.1	14.8	27.2	-0.367298	-0.75543
Solar Radiation (MJ/m2)	8760	0.569111	0.868746	0	0	0.01	0.93	3.52	1.50404	1.126433
Rainfall(mm)	8760	0.148687	1.128193	0	0	0	0	35	14.533232	284.991099
Snowfall (cm)	8760	0.075068	0.436746	0	0	0	0	8.8	8.440801	93.803324

Distributive Statistics



4. Boston Housing

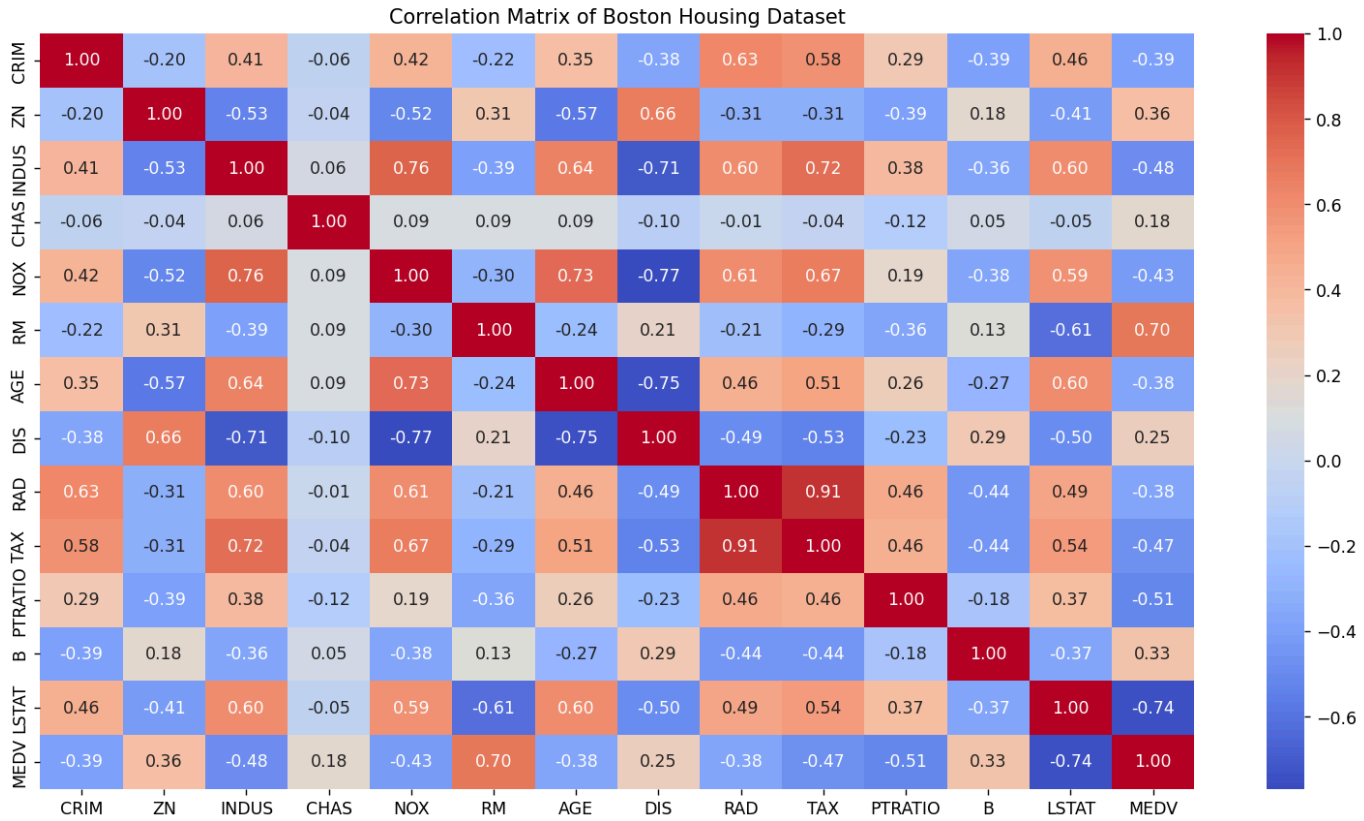
Datatypes

Name	Type	Name	Type
CRIM	float64	DIS	float64
ZN	float64	RAD	int64
INDUS	float64	TAX	float64
CHAS	int64	PTRATIO	float64
NOX	float64	B	float64
RM	float64	LSTAT	float64
AGE	float64	MEDV	float64

Missing Data and Shape

Shape with n/a values: (506, 14)
Shape without n/a values: (506, 14)
n/a values: 0

Correlation Matrix Heatmap



Group Analysis

Average MEDV by CHAS

CHAS	MEDV
0	22.093843
1	28.440000

Average MEDV by RAD

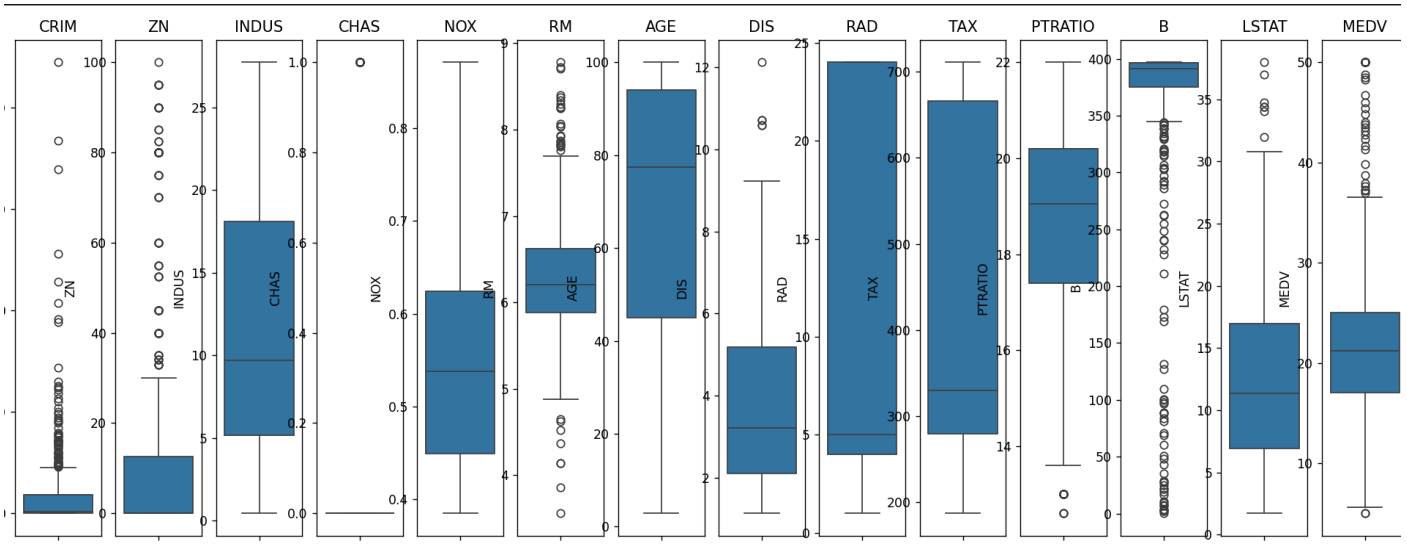
RAD	MEDV
1	24.365000
2	26.833333
3	27.928947
4	21.387273
5	25.706957
6	20.976923
7	27.105882
8	30.358333
24	16.403788

Descriptive Statistics

features	count	mean	std	min	25%	50%	75%	max	skewness	kurtosis
CRIM	506.0	3.613524	8.601545	0.00632	0.082045	0.25651	3.677083	88.9762	5.223149	37.130509
ZN	506.0	11.363636	23.322453	0.00000	0.000000	0.00000	12.500000	100.0000	2.225666	4.031510
INDUS	506.0	11.136779	6.860353	0.46000	5.190000	9.69000	18.100000	27.7400	0.295022	-1.233540

CHAS	506.0	0.069170	0.253994	0.00000	0.000000	0.00000	0.000000	1.0000	3.405904	9.638264
NOX	506.0	0.554695	0.115878	0.38500	0.449000	0.53800	0.624000	0.8710	0.729308	-0.064667
RM	506.0	6.284634	0.702617	3.56100	5.885500	6.20850	6.623500	8.7800	0.403612	1.891500
AGE	506.0	68.574901	28.148861	2.90000	45.025000	77.5000	94.075000	100.0000	-0.598963	-0.967716
DIS	506.0	3.795043	2.105710	1.12960	2.100175	3.20745	5.188425	12.1265	1.011781	0.487941
RAD	506.0	9.549407	8.707259	1.00000	4.000000	5.00000	24.000000	24.0000	1.004815	-0.867232
TAX	506.0	408.237154	168.537116	187.00000	279.000000	330.00000	666.000000	711.0000	0.669956	-1.142408
PTRATIO	506.0	18.455534	2.164946	12.60000	17.400000	19.05000	20.200000	22.0000	-0.802325	-0.285091
B	506.0	356.674032	91.294864	0.32000	375.377500	391.44000	396.225000	396.9000	-2.890374	7.226818
LSTAT	506.0	12.653063	7.141062	1.73000	6.950000	11.3600	16.955000	37.9700	0.906460	0.493240
MEDV	506.0	22.532806	9.197104	5.00000	17.025000	21.2000	25.000000	50.0000	1.108098	1.495197

Distribution Statistics



Quality of Fit

Linear

Auto MPG

	MSE	R2	RMSE
Train	10.71086	0.79015	3.27275
Test	10.71086	0.79015	3.27275
5X Cross-Validation	11.49514	0.81009	3.37577

Forest Fires

	MSE	R2	RMSE
Train	11613.67453	0.01477	107.76676
Test	11613.67453	0.01477	107.76676
5X Cross Validation	4112.06868	-0.19463	55.64507

Seoul Bike Sharing Demand

	MSE	R2	RMSE
Train	194288.2051	0.53369	440.78136
Test	194288.2051	0.53369	440.78136
5X Cross Validation	187603.3027	0.54887	433.0349

Boston Housing

	MSE	R2	RMSE
Train	24.29112	0.66876	4.9286
Test	24.29112	0.66876	4.9286
5X Cross Validation	23.4886	0.71522	4.84279

Ridge

Auto MPG

	MSE	R2	RMSE
Train	10.70278	0.79031	3.27151
Test	10.70278	0.79031	3.27151
Cross Validation	11.49302	0.81013	3.37538

Forest Fires

	MSE	R2	RMSE
Train	11637.64549	0.01273	107.87792
Test	11637.64549	0.01273	107.87792
Cross Validation	4097.53755	-0.16776	55.38703

Seoul Bike Sharing Demand

	MSE	R2	RMSE
Train	194262.2332	0.53375	440.7519
Test	194262.2332	0.53375	440.7519
Cross Validation	187603.5005	0.54887	433.03541

Boston Housing

	MSE	R2	RMSE
Train	24.47719	0.66622	4.94744
Test	24.47719	0.66622	4.94744
Cross Validation	23.71106	0.71222	4.86681

Lasso

Auto MPG

Performance including outliers

	MSE	RSquared	RMSE
In-Sample	10.84749	0.82148	3.29355
Train	11.04147	0.82502	3.32287
Test	10.75401	0.7893	3.27933
5X Validation	11.49538	0.81009	3.37577

Performance excluding outliers - Z Score

	MSE	RSquared	RMSE
In-Sample	10.59414	0.8221	3.25486
Train	10.84212	0.82165	3.29274
Test	10.19649	0.80882	3.19319
5X Validation	11.16959	0.80936	3.33935

Performance excluding outliers - Inter Quartile

	MSE	RSquared	RMSE
In-Sample	10.31223	0.8222	3.21127
Train	10.80482	0.81451	3.28707
Test	8.81355	0.84535	2.96876
5X Validation	10.93944	0.80871	3.30106

Forest Fires

Performance including outliers

	MSE	RSquared	RMSE
In-Sample	3859.08	0.04578	62.12149
Train	1990.089	0.04307	44.61042
Test	11685.56	0.00867	108.0998
5X Validation	4111.496	-0.19373	55.63621

Performance excluding outliers - Z Score

	MSE	RSquared	RMSE
In-Sample	453.2778	0.06339	21.29032
Train	516.6072	0.05447	22.72899
Test	261.3988	-0.15016	16.16783
5X Validation	508.8067	-0.09555	22.06728

Performance excluding outliers - Inter Quartile

	MSE	RSquared	RMSE
In-Sample	12.13703	0.03775	3.48382
Train	12.24535	0.03067	3.49934
Test	13.21939	-0.07467	3.63585
5X Validation	14.5449	-0.17577	3.79941

Seoul Bike Sharing Demand

Performance including outliers

	MSE	RSquared	RMSE
In-Sample	187005.2	0.55044	432.441
Train	185272.8	0.55442	430.4333
Test	194248	0.53378	440.7358
5X Validation	187603.3	0.54887	433.0349

Performance excluding outliers - Z Score

	MSE	RSquared	RMSE
In-Sample	165633.3	0.56597	406.9807
Train	165478.9	0.56784	406.791
Test	166467.1	0.55767	408.0038
5X Validation	166222.4	0.56429	407.6011

Performance excluding outliers - Inter Quartile

	MSE	RSquared	RMSE
In-Sample	156033.1	0.57122	395.0103
Train	155390.7	0.57246	394.1962
Test	159069.4	0.56499	398.8351
5X Validation	156714.3	0.56867	395.8314

Boston Housing

Performance including outliers

	MSE	RSquared	RMSE
In-Sample	21.8952	0.74064	4.67923
Train	22.73981	0.73824	4.76863
Test	25.15559	0.65697	5.01554
5X Validation	23.49094	0.71518	4.84309

Performance excluding outliers - Z Score

	MSE	RSquared	RMSE
In-Sample	10.20754	0.7917	3.19492
Train	10.41864	0.78652	3.22779
Test	10.74615	0.78379	3.27813
5X Validation	11.00931	0.77263	3.28895

Performance excluding outliers - Inter Quartile

	MSE	RSquared	RMSE
In-Sample	4.7584	0.71521	2.18138
Train	5.3524	0.70431	2.31353
Test	3.18889	0.70963	1.78575
5X Validation	5.558	0.66571	2.32556

Transformed

Auto MPG

Including Outliers

Technique	MSE	R2	RMSE
Logarithmic Regression (In-Sample)	8.79549	0.85525	2.96572
Logarithmic Regression (Train)	9.03284	0.85685	3.00547
Logarithmic Regression (Test)	7.6954	0.84923	2.77406
Logarithmic Regression (5X Validation)	0.0136	0.86942	0.11626
Square Root Regression (In-Sample)	10.84748	0.82148	3.29355
Square Root Regression (Train)	9.6417	0.8472	3.10511
Square Root Regression (Test)	8.77465	0.82808	2.9622
Square Root Regression (5X Validation)	11.49514	0.81009	3.37577
Reciprocal Regression (In-Sample)	9.17551	0.84899	3.02911
Reciprocal Regression (Train)	10.11554	0.83969	3.18049
Reciprocal Regression (Test)	7.47045	0.85364	2.73321
Reciprocal Regression (5X Validation)	0.00003	0.87602	0.00579
Yeo-Johnson Regression (In-Sample)	8.99253	0.85201	2.99876
Yeo-Johnson Regression (Train)	9.18992	0.85436	3.03149
Yeo-Johnson Regression (Test)	8.00071	0.84325	2.82855
Yeo-Johnson Regression (5X Validation)	0.0389	0.86294	0.19664
Random Forest (In-Sample)	2.04323	0.96637	1.42942
Random Forest (Train)	1.1296	0.9821	1.06283
Random Forest (Test)	5.66305	0.88905	2.37972
Random Forest (5X Validation)	7.63248	0.8753	2.73176

Excluding Outliers - Z Score

Technique (Including Outliers)	MSE	R2	RMSE
Logarithmic Regression (In-Sample)	8.60586	0.85549	2.93357
Logarithmic Regression (Train)	8.82894	0.85477	2.97135
Logarithmic Regression (Test)	7.25709	0.86393	2.6939
Logarithmic Regression (5X Validation)	0.01321	0.86966	0.11481
Square Root Regression (In-Sample)	10.59413	0.8221	3.25486
Square Root Regression (Train)	9.47738	0.8441	3.07853
Square Root Regression (Test)	8.24672	0.84538	2.87171
Square Root Regression (5X Validation)	11.16989	0.80935	3.3394
Reciprocal Regression (In-Sample)	9.18911	0.84569	3.03136
Reciprocal Regression (Train)	9.25635	0.84774	3.04243
Reciprocal Regression (Test)	6.02293	0.88707	2.45417
Reciprocal Regression (5X Validation)	0.00003	0.87568	0.00576

Yeo-Johnson Regression (In-Sample)	8.84086	0.85154	2.97336
Yeo-Johnson Regression (Train)	9.02431	0.85155	3.00405
Yeo-Johnson Regression (Test)	7.58146	0.85785	2.75345
Yeo-Johnson Regression (5X Validation)	0.04802	0.86146	0.21892
Random Forest (In-Sample)	2.3866	0.95992	1.54486
Random Forest (Train)	1.07251	0.98236	1.03562
Random Forest (Test)	7.64293	0.8567	2.76459
Random Forest (5X Validation)	8.39319	0.85397	2.86884

Excluding Outliers - IQR

Technique (Including Outliers)	MSE	R2	RMSE
Logarithmic Regression (In-Sample)	8.38857	0.85537	2.8963
Logarithmic Regression (Train)	8.5863	0.8526	2.93024
Logarithmic Regression (Test)	7.47503	0.86884	2.73405
Logarithmic Regression (5X Validation)	0.01325	0.86378	0.11485
Square Root Regression (In-Sample)	10.31222	0.8222	3.21126
Square Root Regression (Train)	9.31657	0.84006	3.05231
Square Root Regression (Test)	7.6438	0.86588	2.76474
Square Root Regression (5X Validation)	10.9397	0.80871	3.30111
Reciprocal Regression (In-Sample)	8.82314	0.84788	2.97038
Reciprocal Regression (Train)	8.84068	0.84823	2.97333
Reciprocal Regression (Test)	11.4369	0.79932	3.38185
Reciprocal Regression (5X Validation)	0.00003	0.87008	0.00567
Yeo-Johnson Regression (In-Sample)	8.59721	0.85177	2.9321
Yeo-Johnson Regression (Train)	8.72879	0.85015	2.95445
Yeo-Johnson Regression (Test)	7.4489	0.8693	2.72927
Yeo-Johnson Regression (5X Validation)	0.04502	0.85663	0.2118
Random Forest (In-Sample)	2.02484	0.96509	1.42297
Random Forest (Train)	1.11619	0.98084	1.0565
Random Forest (Test)	5.63519	0.90112	2.37386
Random Forest (5X Validation)	7.48079	0.86995	2.71147

Feature Importance by Random Forest

Outliers	Importance	Z-Score	Importance	IQR	Importance
displacement	0.407415	**displacement**	0.406583	**displacement**	0.361861
horsepower	0.169918	**weight**	0.183966	**weight**	0.21614
cylinders	0.148823	**horsepower**	0.165799	**horsepower**	0.182276
weight	0.139834	**model_year**	0.109298	**model_year**	0.131107
model_year	0.104788	**cylinders**	0.101861	**cylinders**	0.078867
acceleration	0.0248	**acceleration**	0.025738	**acceleration**	0.023404
origin	0.004422	**origin**	0.006756	**origin**	0.006345

Forest Fires

Including Outliers

Technique	MSE	R2	RMSE
Logarithmic Regression (In-Sample)	4130.48147	-0.02133	64.26882
Logarithmic Regression (Train)	2142.32266	-0.03013	46.28523
Logarithmic Regression (Test)	12055.88651	-0.02275	109.7993
Logarithmic Regression (5X Validation)	2.09867	-0.08486	1.44511
Square Root Regression (In-Sample)	3859.07245	0.04578	62.12143
Square Root Regression (Train)	2092.38302	-0.00612	45.74257
Square Root Regression (Test)	11919.3977	-0.01117	109.17599
Square Root Regression (5X Validation)	4112.06868	-0.19463	55.64507
Reciprocal Regression (In-Sample)	4209.27849	-0.04081	64.87895
Reciprocal Regression (Train)	2199.99438	-0.05786	46.9041
Reciprocal Regression (Test)	12174.21723	-0.03278	110.33684
Reciprocal Regression (5X Validation)	2.6995413490827437e+17	-0.08494	519266607.63836
Yeo-Johnson Regression (In-Sample)	4170.5629	-0.03124	64.5799
Yeo-Johnson Regression (Train)	2169.41167	-0.04316	46.57694
Yeo-Johnson Regression (Test)	12504.88511	-0.06084	111.82524
Yeo-Johnson Regression (5X Validation)	0.46485	-0.09749	0.68105
Random Forest (In-Sample)	2744.27978	0.32143	52.38587
Random Forest (Train)	417.75181	0.79912	20.43898
Random Forest (Test)	11983.2803	-0.01659	109.46817
Random Forest (5X Validation)	5005.80994	-1.22527	65.03728

Excluding Outliers - Z Score

Method	MSE	R^2	RMSE
Logarithmic Regression (In-Sample)	512.5734	-0.05913	22.64008
Logarithmic Regression (Train)	579.66766	-0.06095	24.07629
Logarithmic Regression (Test)	234.2028	-0.03049	15.30369
Logarithmic Regression (5X Validation)	1.77478	-0.014	1.33076
Square Root Regression (In-Sample)	453.27181	0.06341	21.29018
Square Root Regression (Train)	555.35671	-0.01645	23.56601
Square Root Regression (Test)	228.16487	-0.00393	15.10513
Square Root Regression (5X Validation)	509.03188	-0.09635	22.07351
Reciprocal Regression (In-Sample)	550.02326	-0.13651	23.45257
Reciprocal Regression (Train)	621.89646	-0.13824	24.93785
Reciprocal Regression (Test)	262.53047	-0.15514	16.20279
Reciprocal Regression (5X Validation)	2.600097448358675e+17	-0.05364	509747646.15471
Yeo-Johnson Regression (In-Sample)	530.15139	-0.09545	23.02502
Yeo-Johnson Regression (Train)	598.23418	-0.09493	24.45883
Yeo-Johnson Regression (Test)	242.47243	-0.06688	15.57153
Yeo-Johnson Regression (5X Validation)	0.42482	-0.0171	0.65154
Random Forest (In-Sample)	182.35145	0.62321	13.50376
Random Forest (Train)	84.4312	0.84547	9.18864
Random Forest (Test)	574.03245	-1.52575	23.95897

Random Forest (5X Validation)	**572.90632**	**-0.3698**	**23.66006**

Excluding Outliers – IQR

Technique	MSE	R^2	RMSE
Logarithmic Regression (In-Sample)	**13.58527**	**-0.07707**	**3.68582**
Logarithmic Regression (Train)	**13.38464**	**-0.05952**	**3.6585**
Logarithmic Regression (Test)	**15.91586**	**-0.29388**	**3.98947**
Logarithmic Regression (5X Validation)	**0.8996**	**-0.17587**	**0.94726**
Square Root Regression (In-Sample)	**12.13689**	**0.03776**	**3.4838**
Square Root Regression (Train)	**13.65537**	**-0.08095**	**3.69532**
Square Root Regression (Test)	**16.61496**	**-0.35071**	**4.07614**
Square Root Regression (5X Validation)	**14.59936**	**-0.18004**	**3.80636**
Reciprocal Regression (In-Sample)	**17.8461**	**-0.41488**	**4.22446**
Reciprocal Regression (Train)	**17.38254**	**-0.37599**	**4.16924**
Reciprocal Regression (Test)	**19.70031**	**-0.60153**	**4.4385**
Reciprocal Regression (5X Validation)	**2.7626537280713123e+17**	**-0.14939**	**525386874.28687**
Yeo-Johnson Regression (In-Sample)	**14.94084**	**-0.18454**	**3.86534**
Yeo-Johnson Regression (Train)	**14.8387**	**-0.17462**	**3.8521**
Yeo-Johnson Regression (Test)	**17.25788**	**-0.40298**	**4.15426**
Yeo-Johnson Regression (5X Validation)	**0.24252**	**-0.16913**	**0.49219**
Random Forest (In-Sample)	**4.62221**	**0.63354**	**2.14993**

Random Forest (Train)	**2.24398**	**0.82237**	**1.49799**
Random Forest (Test)	**14.1351**	**-0.14911**	**3.75967**
Random Forest (5X Validation)	**14.87006**	**-0.20507**	**3.8434**

Feature Importance by Random Forest

Outliers	Importance	Z-Score	Importance	IQR	Importance
DMC	0.1694071	**Y**	0.1268003	**temp**	0.160936
temp	0.1672118	**wind**	0.1225544	**DMC**	0.152375
RH	0.1031345	**temp**	0.1104956	**DC**	0.143829
wind	0.08915285	**RH**	0.09638183	**RH**	0.123305
Y	0.07499659	**X**	0.09459988	**X**	0.085732
FFMC	0.0681802	**day_sat**	0.09389508	**wind**	0.076619
X	0.05915375	**DMC**	0.0765702	**ISI**	0.068197
ISI	0.05447728	**ISI**	0.07415865	**FFMC**	0.06323
DC	0.05319513	**DC**	0.06425293	**Y**	0.042156
day_thu	0.03664583	**FFMC**	0.04800677	**day_thu**	0.020648
day_sat	0.02748757	**month_sep**	0.02061919	**day_sat**	0.017743
day_mon	0.02707548	**day_sun**	0.01930376	**day_sun**	0.011602
month_jul	0.0208649	**month_aug**	0.01083513	**day_tue**	0.009875
day_sun	0.01346987	**month_jul**	0.007836371	**day_wed**	0.004913
month_aug	0.01114579	**day_wed**	0.006012869	**day_mon**	0.0049
day_wed	0.006549027	**month_may**	0.005394649	**month_sep**	0.004107
month_sep	0.006060789	**month_oct**	0.004655547	**month_aug**	0.003708
day_tue	0.004231065	**day_mon**	0.004145851	**month_oct**	0.002406
month_may	0.002547817	**day_tue**	0.004029233	**month_jul**	0.002023
month_jun	0.001304156	**month_mar**	0.003007553	**month_jun**	0.001696
month_oct	0.001002762	**day_thu**	0.00271643	**rain**	0.0
rain	0.0008986776	**month_jun**	0.001843978	**month_nov**	0.0
month_dec	0.0006832938	**month_dec**	0.0009765764	**month_may**	0.0
month_mar	0.000580944	**month_feb**	0.0009041777	**month_dec**	0.0
month_feb	0.0005421003	**rain**	2.128452e-06	**month_feb**	0.0

month_jan	8.185991e-07	**month_nov**	7.277916e-07	**month_mar**	0.0
month_nov	0.0	**month_jan**	2.709299e-07	**month_jan**	0.0

Seoul Bike Sharing Demand

Including Outliers

Technique	MSE	R^2	RMSE
Logarithmic Regression (In-Sample)	206,857.11	0.50272	454.81546
Logarithmic Regression (Train)	202,734.04	0.51243	450.25997
Logarithmic Regression (Test)	206,819.82	0.50361	454.77447
Logarithmic Regression (5X Validation)	0.52405	0.78905	0.72383
Square Root Regression (In-Sample)	187,005.22	0.55044	432.441
Square Root Regression (Train)	171,967.73	0.58642	414.68992
Square Root Regression (Test)	181,108.93	0.56532	425.56895
Square Root Regression (5X Validation)	187,603.30	0.54887	433.0349
Reciprocal Regression (In-Sample)	525,784,936.83	-1262.98436	22,930.00
Reciprocal Regression (Train)	1,949,959,711.99	-4688.58893	44,158.35
Reciprocal Regression (Test)	422,624,754.90	-1013.34914	20,557.84
Reciprocal Regression (5X Validation)	0.00023	1	0.01501
Yeo-Johnson Regression (In-Sample)	177,519.91	0.57324	421.33112
Yeo-Johnson Regression (Train)	175,263.43	0.5785	418.64475
Yeo-Johnson Regression (Test)	182,781.83	0.5613	427.52992
Yeo-Johnson Regression (5X Validation)	31.55941	0.68067	5.61675
Random Forest (In-Sample)	17,355.47	0.95828	131.74016
Random Forest (Train)	7,085.55	0.98296	84.17568
Random Forest (Test)	58,435.17	0.85975	241.73368
Random Forest (5X Validation)	53,805.30	0.87062	231.83811

Excluding Outliers - Z Score

Technique	MSE	R^2	RMSE
Logarithmic Regression (In-Sample)	172,312.86	0.54847	415.10584
Logarithmic Regression (Train)	173,454.45	0.54701	416.47863
Logarithmic Regression (Test)	168,224.61	0.553	410.15194
Logarithmic Regression (5X Validation)	0.44641	0.81686	0.66761
Square Root Regression (In-Sample)	165,633.33	0.56597	406.98074
Square Root Regression (Train)	152,962.70	0.60053	391.10446
Square Root Regression (Test)	151,737.95	0.59681	389.53556
Square Root Regression (5X Validation)	166,222.45	0.56429	407.60112

Reciprocal Regression (In-Sample)	64,002,752.68	-166.71482	8000.17204
Reciprocal Regression (Train)	161,135,529.36	-419.81748	12,693.92
Reciprocal Regression (Test)	19,018,082.47	-49.53401	4,360.97
Reciprocal Regression (5X Validation)	0.00016	1	0.01187
Yeo-Johnson Regression (In-Sample)	155,046.11	0.59371	393.75894
Yeo-Johnson Regression (Train)	155,027.78	0.59513	393.73567
Yeo-Johnson Regression (Test)	152,997.33	0.59346	391.14873
Yeo-Johnson Regression (5X Validation)	35.38454	0.69676	5.94686
Random Forest (In-Sample)	15,954.48	0.95819	126.31103
Random Forest (Train)	7,003.70	0.98171	83.68813
Random Forest (Test)	51,746.75	0.8625	227.47913
Random Forest (5X Validation)	52,105.48	0.86345	227.97348

Excluding Outliers - IQR

Technique	MSE	R^2	RMSE
Logarithmic Regression (In-Sample)	161,361.16	0.55658	401.69785
Logarithmic Regression (Train)	160,411.45	0.55864	400.51399
Logarithmic Regression (Test)	178,091.71	0.51297	422.00914
Logarithmic Regression (5X Validation)	0.41038	0.82882	0.63977
Square Root Regression (In-Sample)	156,033.13	0.57122	395.0103
Square Root Regression (Train)	143,548.04	0.60504	378.87734
Square Root Regression (Test)	150,372.60	0.58877	387.77906
Square Root Regression (5X Validation)	156,714.32	0.56867	395.83139
Reciprocal Regression (In-Sample)	290,483,070.31	-797.24182	17,043.56
Reciprocal Regression (Train)	392,174,634.16	-1078.03454	19,803.40
Reciprocal Regression (Test)	1,233,987,729.78	-3373.61685	35,128.16
Reciprocal Regression (5X Validation)	0.00009	1	0.00883
Yeo-Johnson Regression (In-Sample)	145,723.35	0.59956	381.73728
Yeo-Johnson Regression (Train)	144,736.75	0.60177	380.44284
Yeo-Johnson Regression (Test)	152,801.40	0.58213	390.8982
Yeo-Johnson Regression (5X Validation)	40.88786	0.69084	6.39325
Random Forest (In-Sample)	15,745.80	0.95673	125.48228
Random Forest (Train)	7,094.73	0.98048	84.23022
Random Forest (Test)	50,350.09	0.86231	224.38826
Random Forest (5X Validation)	50,758.63	0.86032	225.25683

Boston Housing

Including Outliers

Technique	MSE	R^2	RMSE
Logarithmic Regression (In-Sample)	18.59032	0.77979	4.31165
Logarithmic Regression (Train)	18.32633	0.78905	4.28093
Logarithmic Regression (Test)	17.30027	0.76409	4.15936
Logarithmic Regression (5X Validation)	0.03409	0.7678	0.18462
Square Root Regression (In-Sample)	21.89483	0.74064	4.67919
Square Root Regression (Train)	19.13472	0.77974	4.37433
Square Root Regression (Test)	18.82078	0.74335	4.33829
Square Root Regression (5X Validation)	23.4886	0.71522	4.84279
Reciprocal Regression (In-Sample)	20.73025	0.75444	4.55305
Reciprocal Regression (Train)	20.57254	0.76319	4.5357
Reciprocal Regression (Test)	18.88658	0.74246	4.34587
Reciprocal Regression (5X Validation)	0.0002	0.67806	0.01398
Yeo-Johnson Regression (In-Sample)	18.77313	0.77762	4.3328
Yeo-Johnson Regression (Train)	18.53051	0.7867	4.30471
Yeo-Johnson Regression (Test)	17.69277	0.75874	4.20628
Yeo-Johnson Regression (5X Validation)	0.09736	0.76587	0.31201
Random Forest (In-Sample)	3.27191	0.96124	1.80884
Random Forest (Train)	1.94937	0.97756	1.3962
Random Forest (Test)	8.5102	0.88395	2.91722
Random Forest (5X Validation)	10.80462	0.87056	3.25515

Excluding Outliers - Z Score

Technique	MSE	R^2	RMSE
Logarithmic Regression (In-Sample)	8.73262	0.8218	2.9551
Logarithmic Regression (Train)	8.76663	0.82037	2.96085
Logarithmic Regression (Test)	7.7513	0.84404	2.78411
Logarithmic Regression (5X Validation)	0.02279	0.77532	0.14798
Square Root Regression (In-Sample)	10.20716	0.79171	3.19486
Square Root Regression (Train)	9.13453	0.81283	3.02234
Square Root Regression (Test)	8.63013	0.82636	2.93771
Square Root Regression (5X Validation)	11.00581	0.7727	3.2883

Reciprocal Regression (In-Sample)	10.29787	0.78986	3.20903
Reciprocal Regression (Train)	10.60316	0.78274	3.25625
Reciprocal Regression (Test)	12.68893	0.7447	3.56215
Reciprocal Regression (5X Validation)	0.00014	0.67471	0.01125
Yeo-Johnson Regression (In-Sample)	9.05145	0.81529	3.00856
Yeo-Johnson Regression (Train)	9.15589	0.81239	3.02587
Yeo-Johnson Regression (Test)	8.66372	0.82569	2.94342
Yeo-Johnson Regression (5X Validation)	0.27016	0.78258	0.51247
Random Forest (In-Sample)	2.41838	0.95065	1.55511
Random Forest (Train)	1.10016	0.97746	1.04888
Random Forest (Test)	7.67498	0.84558	2.77037
Random Forest (5X Validation)	7.98042	0.83341	2.8131

Excluding Outliers - IQR

Technique	MSE	R^2	RMSE
Logarithmic Regression (In-Sample)	4.70766	0.71825	2.16971
Logarithmic Regression (Train)	5.17419	0.71415	2.27468
Logarithmic Regression (Test)	2.92426	0.73373	1.71005
Logarithmic Regression (5X Validation)	0.01053	0.6569	0.10178
Square Root Regression (In-Sample)	4.75707	0.71529	2.18107
Square Root Regression (Train)	5.14101	0.71599	2.26738
Square Root Regression (Test)	3.0734	0.72015	1.75311
Square Root Regression (5X Validation)	5.55515	0.66583	2.32526
Reciprocal Regression (In-Sample)	4.98561	0.70161	2.23285
Reciprocal Regression (Train)	5.6024	0.6905	2.36694
Reciprocal Regression (Test)	2.90066	0.73588	1.70313
Reciprocal Regression (5X Validation)	0.00003	0.6157	0.00538
Yeo-Johnson Regression (In-Sample)	4.69925	0.71875	2.16778
Yeo-Johnson Regression (Train)	5.1706	0.71435	2.27389
Yeo-Johnson Regression (Test)	2.92989	0.73322	1.71169
Yeo-Johnson Regression (5X Validation)	0.02492	0.65973	0.15649
Random Forest (In-Sample)	1.58292	0.90526	1.25814
Random Forest (Train)	0.88394	0.95117	0.94018
Random Forest (Test)	4.36259	0.60276	2.08868
Random Forest (5X Validation)	5.83846	0.64417	2.39711

Symbolic

Auto MPG

In-Sample:

In-Sample Mean Squared Error: 11.189024963623158
In-Sample R-squared: 0.8158571313528716
In-Sample Adjusted R-squared: 0.812500360309825
In-Sample Mean Absolute Error: 2.353617718343498

Train Test Split(80-20):

Validation Mean Squared Error: 12.136821723043147
Validation R-squared: 0.762212322966489
Validation Adjusted R-squared: 0.7387684674843118
Validation Mean Absolute Error: 2.4698216424090864

5X Cross Validation:

5x Cross-Validation Mean Squared Error: 11.829055107927978
5x Cross-Validation R2 Score: 0.801020894

Forest Fires

In-Sample:

In-Sample Mean Squared Error: 4185.389555705995
In-Sample R-squared: -0.03490507010978616
In-Sample Adjusted R-squared: -0.0920470678459091
In-Sample Mean Absolute Error: 12.81634429400387

Train Test Split(80-20):

Validation Mean Squared Error: 12172.81992307692
Validation R-squared: -0.03266589522453578
Validation Adjusted R-squared: -0.3995340422121998
Validation Mean Absolute Error: 19.63923076923077

5X Cross Validation:

5x Cross-Validation Mean Squared Error: 4192.789134191329
5x Cross-Validation R2 Score: −0.074843213.

Seoul Bike Sharing Demand

In-Sample:

In-Sample Mean Squared Error: 160908.3946181571
In-Sample R-squared: 0.6131770233681545
In-Sample Adjusted R-squared: 0.6125577527366113
In-Sample Mean Absolute Error: 268.87295720647523

Train Test Split(80-20):

Validation Mean Squared Error: 184405.1696326787
Validation R-squared: 0.5574058927138987
Validation Adjusted R-squared: 0.5538386402659969
Validation Mean Absolute Error: 286.3442868160291

5X Cross Validation:

5x Cross-Validation Mean Squared Error: 167078.8076922553
5x Cross-Validation R2 Score: 0.5982950184899063

Boston Housing

In-Sample:

In-Sample Mean Squared Error: 25.839619066954363
In-Sample R-squared: 0.6939142984931794
In-Sample Adjusted R-squared: 0.6858266681688122
In-Sample Mean Absolute Error: 3.5066574723639965

Train Test Split(80-20):

Validation Mean Squared Error: 33.00715822561788
Validation R-squared: 0.5499051487162835
Validation Adjusted R-squared: 0.4834138638675527
Validation Mean Absolute Error: 4.1133147595864425

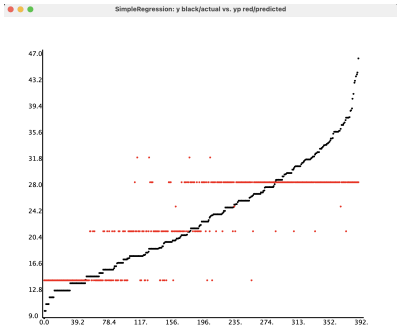
5X Cross Validation:

5x Cross-Validation Mean Squared Error: 39.417021579726295
5x Cross-Validation R2 Score: 0.528097465439966

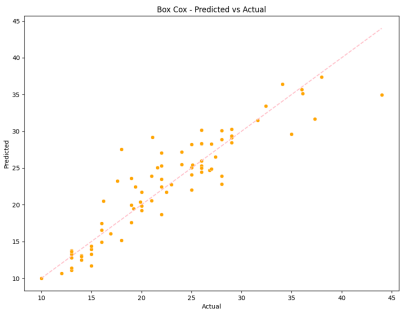
Statistical Summaries and Plots

Auto MPG

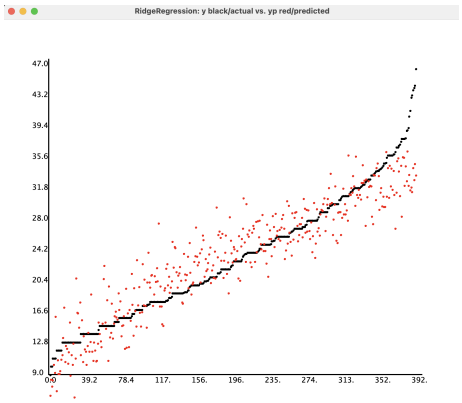
Linear



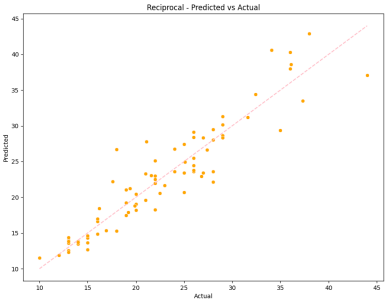
Box Cox



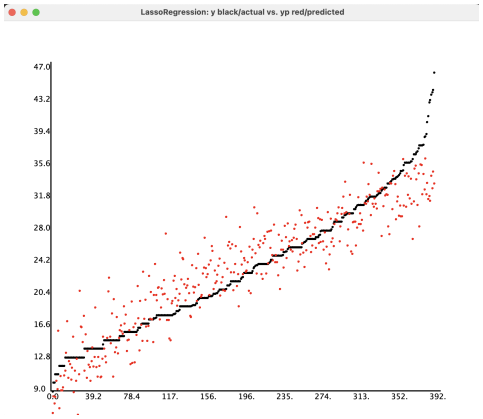
Ridge



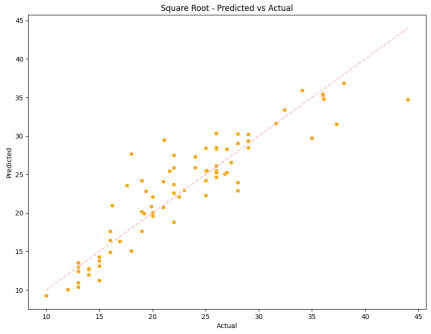
Reciprocal



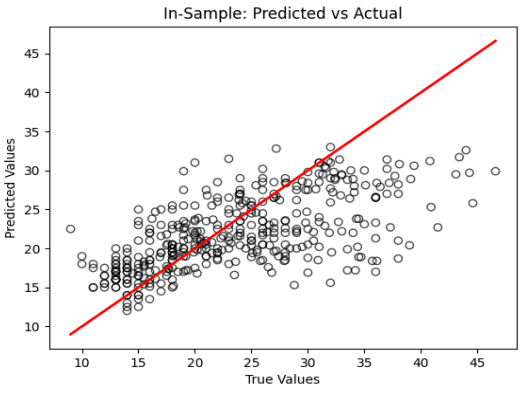
Lasso



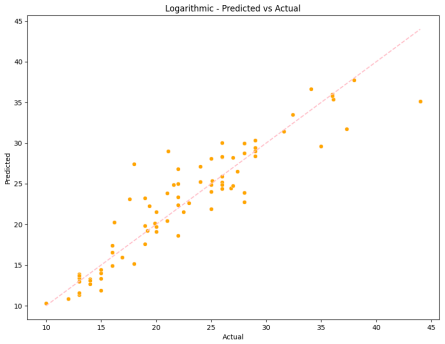
Square Root



Symbolic

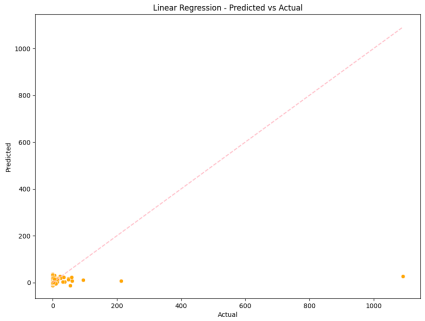


Logarithmic

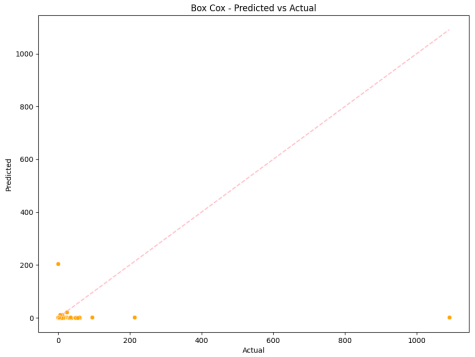


Forest Fire

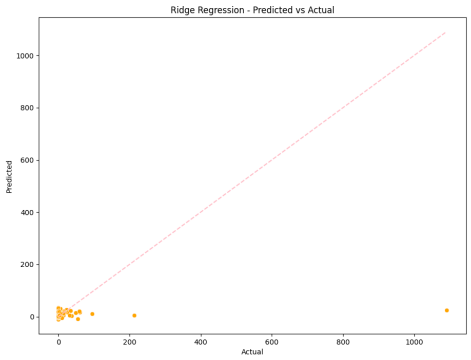
Linear



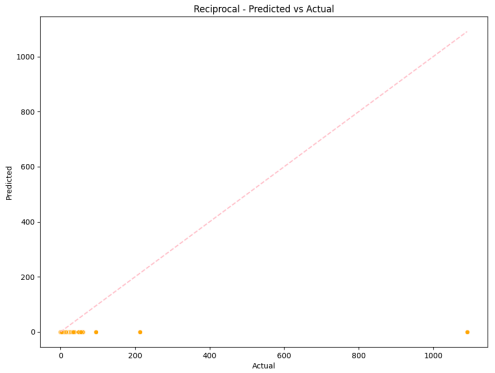
Box Cox



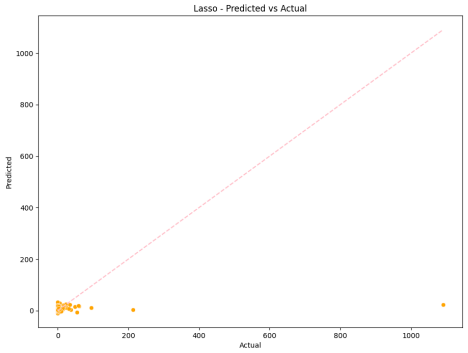
Ridge



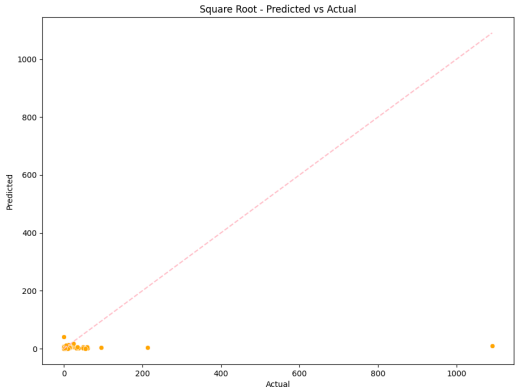
Reciprocal



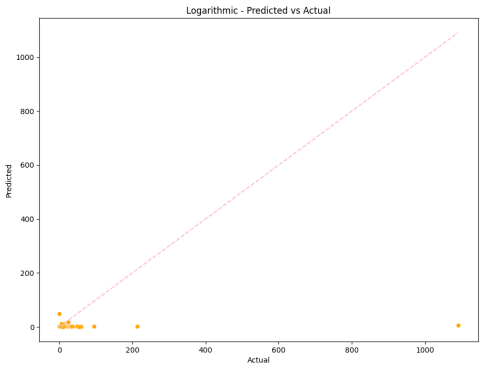
Lasso



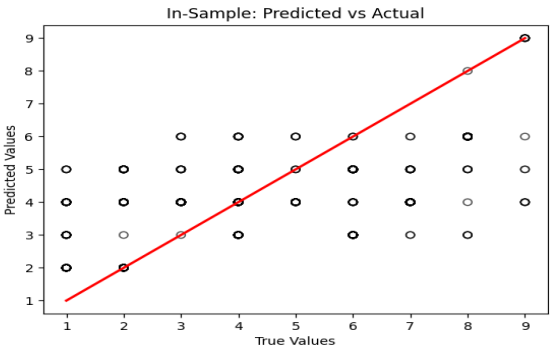
Square Root



Logarithmic

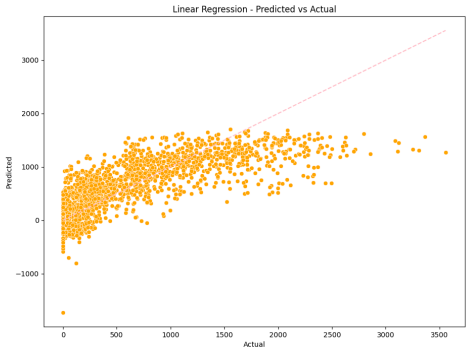


Symbolic

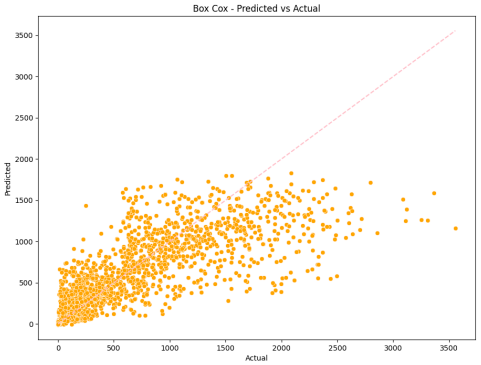


Seoul Bike Demand

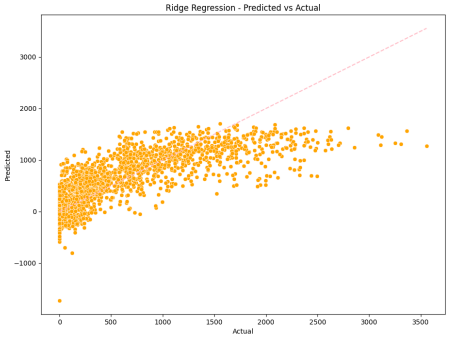
Linear



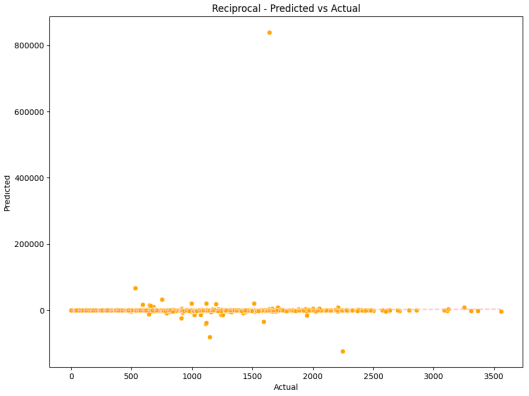
Box Cox



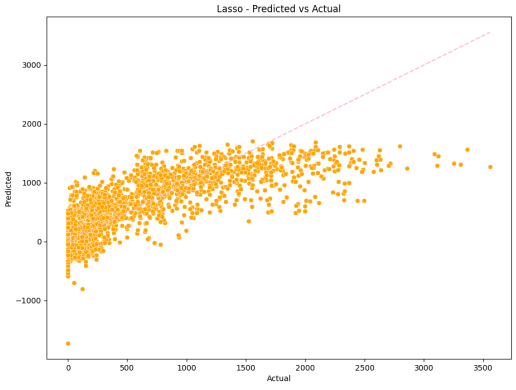
Ridge



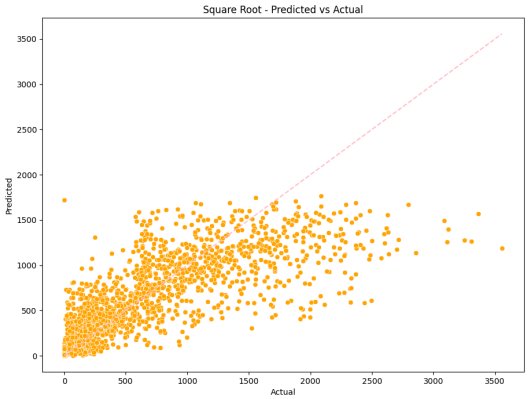
Reciprocal



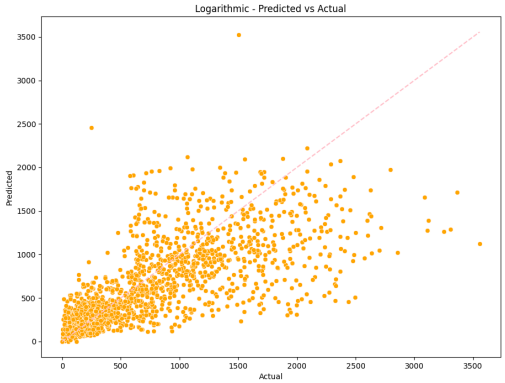
Lasso



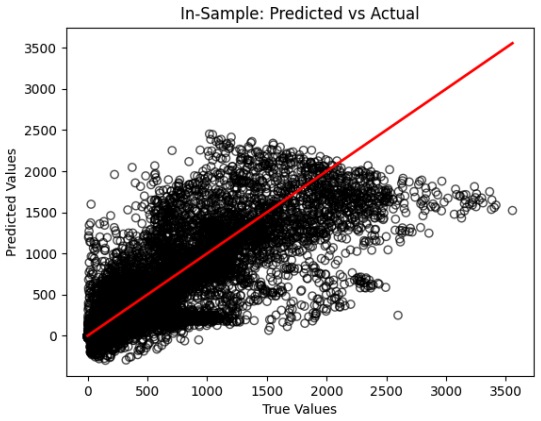
Square Root



Logarithmic

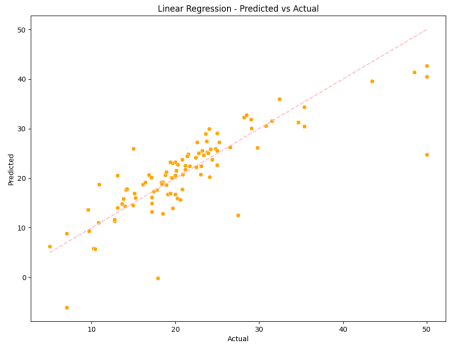


Symbolic

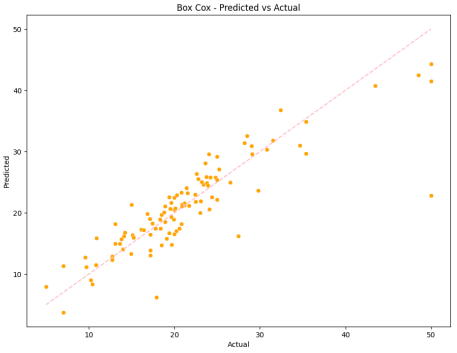


Boston

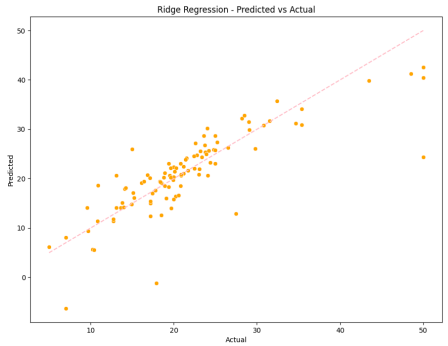
Linear



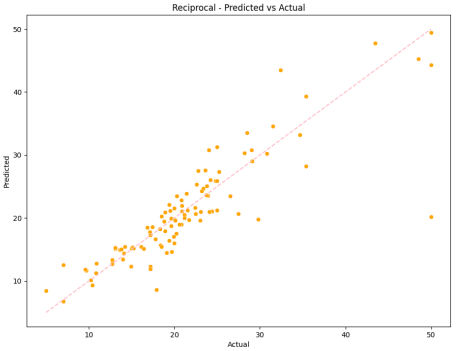
Box Cox



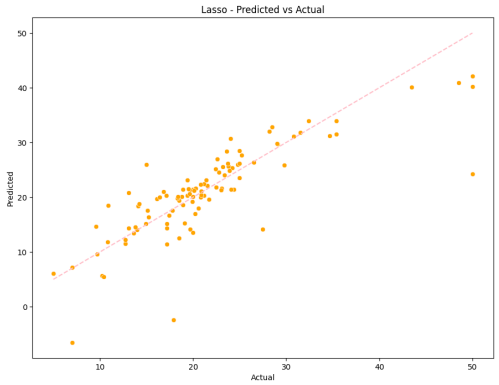
Ridge



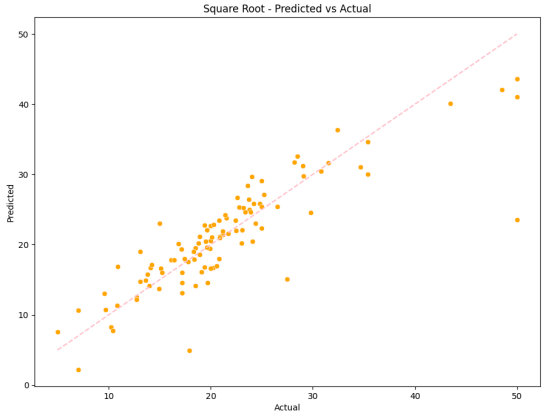
Reciprocal



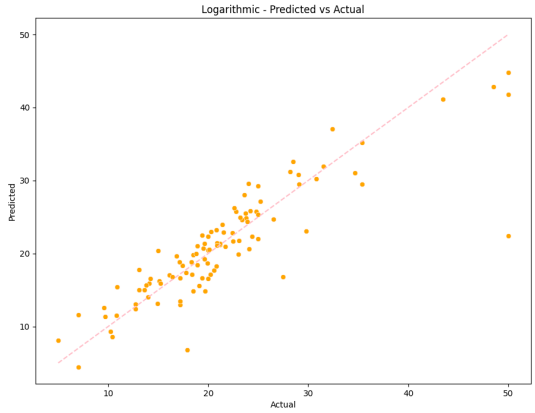
Lasso



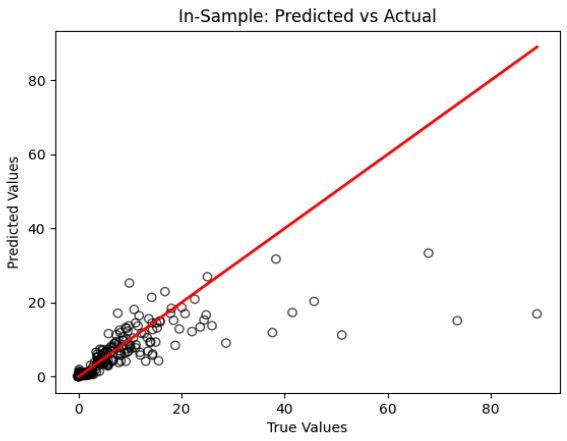
Square Root



Logarithmic



Symbolic



Feature Selection

Auto MPG Selected Features
Forward: [weight, model_year, origin, displacement, horsepower, cylinders, acceleration]
Backward: [cylinders, displacement, horsepower, weight, acceleration, model_year, origin]
Stepwise: [weight]

Forest Fire Selected Features
Forward: [temp, day_sat, X, month_sep, wind, month_dec, ISI, DMC, DC, month_oct, month_aug, month_jul, day_thu, month_mar, RH, day_tue, rain, day_mon, day_sun, day_wed, month_jan, month_may, month_feb, month_jun, FFMC, month_nov, Y]
Backward: [temp, day_sat]
Stepwise: [temp]

Seoul Bike Share Demand Selected Features
Forward: [Temperature(C), Hour, Functioning Day_Yes, Humidity(%), Seasons_Winter, Rainfall(mm), Solar Radiation (MJ/m2), Seasons_Spring, Seasons_Summer, Holiday_No Holiday, Wind speed (m/s), Dew point temperature(C), Snowfall (cm), Visibility (10m)]
Backward: [Hour, Temperature(C), Humidity(%), Wind speed (m/s), Visibility (10m), Dew point temperature(C), Solar Radiation (MJ/m2), Rainfall(mm), Snowfall (cm), Seasons_Spring, Seasons_Summer, Seasons_Winter, Holiday_No Holiday, Functioning Day_Yes]
Stepwise: [Seasons_Winter]

Boston Housing Selected Features
Forward: [LSTAT, RM, PTRATIO, DIS, NOX, CHAS, B, ZN, CRIM, RAD, TAX, INDUS, AGE]
Backward: [CRIM, ZN, INDUS, CHAS, NOX, RM, AGE, DIS, RAD, TAX, PTRATIO, B, LSTAT]
Stepwise: [LSTAT]

Auto MPG				**Forest Fire**		
F-Selection	**R Squared**	**R Squared Adj**		**F-Selection**	**R Squared**	**R Squared Adj**
Forward	0.821478	0.818224		**Forward**	0.045782	-0.006905
Backward	0.821478	0.818224		**Backward**	0.016714	0.012888
Stepwise	0.69263	0.691842		**Stepwise**	0.009573	0.00765
Seoul				**Boston**		
Forward	0.55044	0.549721		**Forward**	0.740643	0.73379
Backward	0.55044	0.549721		**Backward**	0.740643	0.73379
Stepwise	0.180561	0.180468		**Stepwise**	0.544146	0.543242

Discussion of Results

Auto MPG:

- Brands like Nissan and Honda show higher average MPG, indicating more fuel-efficient vehicles.
- Cars with 4 cylinders have highest average MPG (29.28), while those with 8 cylinders have the lowest (14.96)
- A steady increase in MPG is observed over model years, reflecting improvements in fuel efficiency over time.
- Horsepower and weight exhibit high variability, while skewness and kurtosis indicate some non-normality in distributions.
- In-Sample: Logarithmic Regression has the highest R^2 (0.85525) indicating that it explains more variance of the target variable within the sample.
- Train: Logarithmic Regression with R^2 of 0.85685.
- Test: Logarithmic Regression with R^2 of 0.84923.
- Cross-Validation: Logarithmic Regression with R^2 of 0.86942.

Forest Fires:

- May and September have the highest average burned area, suggesting these months are more prone to larger fires.
- Saturday has the highest average burned area, potentially reflecting human activities affecting fire occurrence or reporting bias.
- Higher average burned area is noted for zero rainfall, indicating dry conditions as a key factor.
- Temperature, wind speed, and drought code (DC) influence the burned area more, based on feature importance.
- In-Sample: Square Root Regression has the highest R^2 of 0.04578.
- Train: Square Root Regression with R^2 of 0.04578.
- Test: Yeo-Johnson Regression with the least negative R^2 (-0.03278), indicating slightly better performance among weak models.
- Cross-Validation: Square Root Regression with the highest negative R^2 (-0.19463).

Seoul Bike Sharing Demand:

- Demand peaks during summer, while winter has the lowest demand, reflecting expected seasonal variation.
- Dew Point Temperature and Temperature are highly correlated.
- Significant increases in bike rentals are observed during evening hours (5 PM to 7 PM), likely due to commuting patterns.
- Average rentals are lower on holidays, indicating reduced commuting and business activities.
- In-Sample: Yeo-Johnson Regression has the highest R^2 of 0.57324.
- Train: Yeo-Johnson Regression with R^2 of 0.5785.
- Test: Yeo-Johnson Regression with R^2 of 0.5613.
- Cross-Validation: Yeo-Johnson Regression with R^2 of 0.68067.

Boston Housing:

- Homes near the Charles River (CHAS = 1) have significantly higher average prices
- Higher access to radial highways (RAD = 8) is associated with higher median housing prices.
- In-Sample: Yeo-Johnson Regression has the highest R^2 of 0.77762.
- Train: Yeo-Johnson Regression with R^2 of 0.7867.
- Test: Logarithmic Regression with R^2 of 0.76409.
- Cross-Validation: Logarithmic Regression with R^2 of 0.7678.

Misc:

- **Higher R^2 values indicate better fit and more explained variance by the model.**
- **For datasets with lower or negative R^2 (Forest Fires), it suggests that the models are not capturing the data patterns effectively.**
- **Cross-validation results show consistency and reliability across different data splits, with high R^2 values suggesting robust model performance.**